Kai Carstensen und Klaus Wohlrabe und Christina Ziegler:

Predictive Ability of Business Cycle Indicators under Test: A Case Study for the Euro Area Industrial Production

Munich Discussion Paper No. 2010-16

Department of Economics
University of Munich

Volkswirtschaftliche Fakultät
Ludwig-Maximilians-Universität München

Online at http://epub.ub.uni-muenchen.de/11442/

Predictive Ability of Business Cycle Indicators under Test: A Case Study for the Euro Area Industrial Production

Kai Carstensen

Ifo Institute for Economic Research and University of Munich, Germany.

Klaus Wohlrabe[1]

Ifo Institute for Economic Research, Germany.

Christina Ziegler

Ifo Institute for Economic Research and University of Leipzig, Germany.

This version: February 2010

Preliminary — Please do not quote

[1]Corresponding author. E-mail: wohlrabe@ifo.de.

Abstract

In this paper we assess the information content of seven widely cited early indicators for the euro area with respect to forecasting area-wide industrial production. To this end, we use various tests that are designed to compare competing forecast models. In addition to the standard Diebold-Mariano test, we employ tests that account for specific problems typically encountered in forecast exercises. Specifically, we pay attention to nested model structures, we alleviate the problem of data snooping arising from multiple pairwise testing, and we analyze the structural stability in the relative forecast performance of one indicator compared to a benchmark model. Moreover, we consider loss functions that overweight forecast errors in booms and recessions to check whether a specific indicator that appears to be a good choice on average is also preferable in times of economic stress. We find that on average three indicators have superior forecast ability, namely the EuroCoin indicator, the OECD composite leading indicator, and the FAZ-Euro indicator published by the Frankfurter Allgemeine Zeitung. If one is interested in one-month forecasts only, the business climate indicator of the European Commission yields the smallest errors. However, the results are not completely invariant against the choice of the loss function. Moreover, rolling local tests reveal that the indicators are particularly useful in times of unusual changes in industrial production while the simple autoregressive benchmark is difficult to beat during time of average production growth.

JEL-numbers: C32, C53, E32.

Keywords: weighted loss, leading indicators, euro area, forecasting.

1 Introduction

The euro area is a rather new subject in the literature on macroeconomic forecasting. However, it is all the more interesting, especially because the European Central Bank conducts its monetary policy explicitly with a view to the euro area as a whole. The forward-looking elements of this policy requires to generate accurate forecasts of inflation and economic activity. In this paper, we consider the latter, concentrating on euro area industrial production which is the most timely "hard indicator" of aggregate output that is available. Specifically, we assess whether several popular "soft indicators" reveal early information that helps to improve the accuracy of industrial production forecasts.

In standard empirical out-of-sample forecasting exercises the performance of leading indicators is often measured by the (root) mean squared error which is derived from a symmetric quadratic loss function. Furthermore, in order uncover significantly forecasting differences between pairs of indicators, typically the popular Diebold-Mariano test is employed.

In line with the recent literature, we challenge this "standard assessment approach" in several ways. First, we allow for a flexible weighting scheme of the forecasting errors in the relevant loss function. This can be more satisfactory in situations where some observations are more important than others, as argued by van Dijk and Franses (2003). The flexible weighting scheme allows to judge the predictive ability of leading indicators during booms or recessions which might be particularly important times for monetary policy decisions and, thus, accurate forecasts, the recent financial and economic crisis being an impressive example. To take these issues into account, we include a weighted loss function into the standard Diebold-Mariano type tests.

Second, we pay attention to the the aspect of nested models in forecast comparisons. Starting with Clark and McCracken (2001) this aspect has been increasingly discussed in the literature. The basic idea is that the comparison of, say, an indicator model with a nested benchmark model (that does not include the indicator) has to take into account the estimation uncertainty associated with estimating the additional parameters for the indicators. Neglecting this uncertainty gives rise to a bias in favor of the benchmark model. For example, in such a situation the Diebold-Mariano test would signal too often that the indicator model is not able to improve upon the benchmark. Specifically, we employ the recently proposed test by Clark and West (2007) to account for this issue.

Third, we note that our forecast comparison—like almost all work in this field—does not literally contrast one model with a single competitor which is the setting the standard pairwise tests such as the one proposed by Diebold and Mariano (1995) are designed for. Instead, we aim at finding the most promising indicators from a possibly large set of can-

didates. In such a situation, a few pairwise tests can signal dominance of one indicator over the other simply by chance, much like repeated draws from, say, the standard normal distribution will yield from time to time values that exceed conventional critical values and lead to the rejection of the mean zero hypothesis. To account for this data snooping problem we apply the test for superior predictive ability (SPA) proposed by Hansen (2005) and based on the seminal paper by White (2000).

Finally, we take a first look at the stability issue of forecast dominance. As argued by Giacomini and Rossi (2008) the relative forecast performance of one indicator to another may change over time, possibly due to structural instabilities, e.g., as the consequence of booms or recessions. A practitioner would of course prefer an indicator that has at least in past shown stable dominance over its competitors. To this end, we implement the fluctuation test proposed by Giacomini and Rossi (2008) which is based on a series of local Diebold-Mariano tests. To the best of our knowledge, we are the first who allow for weighted loss differentials within this framework to assess the forecasting stability also for booms and recessions.

The remainder of the paper is structured as follows. In Section 2, we briefly overview the related literature. In Section 3, we discuss the weighted loss function we use to compare to forecast models before we outline in Section 4 the various forecast accuracy tests we employ. The setup of our out-of-sample forecast exercise is described in Section 5 and the results are presented in Section 6. Section 7 summarizes and concludes.

2 Related Literature

As the euro area is a rather new entity, it is has become only recently a topic in the field of macroeconomic forecasting. Accordingly, there are only few directly related papers available. While we study point forecasts, most of the work done on the euro area focuses on turning point prediction for industrial production, or point forecasts for gross domestic product and inflation. Only the study by Bodo et al. (2000) uses one of the indicators we consider, namely the European Economic Sentiment indicator. Therefore, we are among the first who assess the point forecasting ability of leading indicators for the euro area.

Bodo et al. (2000) provide one of the first studies to forecast euro area industrial production. Besides univariate and vector autoregressive models referring to the four largest euro area countries, the authors employ a two-country vector autoregressive model for the euro area and the US. They study whether the inclusion of survey-based business climate indicator published by the European Commission helps to improve the forecasts. Employing the modified Diebold-Mariano test, they find that the benchmark ARIMA model is

outperformed by the two-country model with the survey indicator.

Marcellino et al. (2003) forecast quarterly euro area macroeconomic time series, among them industrial production, using a dynamic factor model framework with country-specific data. They find, based on a number of different model specifications, that country-specific information matters, albeit without testing for significant differences in predictive ability. Forni et al. (2003) show in a dynamic factor framework that including financial variables does not improve forecast accuracy for euro area industrial production. Marcellino (2008) provides evidence that artificial neural networks perform on average better than simple linear models without indicators.

Using different forecast targets, there are quite a few papers that apply the newly developed tests of forecast accuracy discussed above. However, they typically focus on exchange rate and financial forecasting. As an exception, Milas and Rothman (2008) use weighted loss differentials as proposed by van Dijk and Franses (2003) to assess macroeconomic forecasting performance. They use smooth transition vector error-correction models in a simulated out-of-sample forecasting experiment for the unemployment rates in the U.S., the U.K., Canada, and Japan. They find that the forecast performance of the models can differ between booms and recessions. Caggiano et al. (2009) use the test proposed by Clark and West (2007) to account for nested model structures when comparing forecast models for the euro area and other countries. The aspect of data snooping has recently been taken into account by Clark and McCracken (2009) who compare a very large set of forecasting models for U.S. macroeconomic variables. The fluctuation test is used in Fichtner et al. (2009) to assess the stability in the predictive ability of the OECD composite leading indicator for industrial production in 11 OECD countries. It is also used by Rossi and Sekhposyan (2010) to check whether the forecasting performance of various economic models for US output growth and inflation has changed over time. They find that during the Great Moderation many forecasting models became essentially useless.

3 Weighted Loss Functions

The standard period-t loss function used in most of the forecast evaluation literature ist the squared forecast error

$$\mathcal{L}_{i,t} = e_{i,t}^2, \tag{1}$$

where $e_{i,t} = y_t - y_{i,t}^f$ is the forecast error of model i, y_t is the realization of the target variable, $y_{i,t}^f$ is the value predicted by model i. While theoretical results are available for quite general loss functions, see, e.g., Diebold and Mariano (1995), the applied literature concentrates on the quadratic loss function. Comparing the average loss difference of two competing

models 1 and 2 then means to compute their mean squared forecast errors (MSFE)

$$\text{MSFE}_i = \frac{1}{P}\sum_{t=T+1}^{T+P} e_{i,t}^2, \quad i = 1,2, \tag{2}$$

over the forecast period $T+1$ to $T+P$ and choose the model with the smaller MSFE. However, one can think of many occasions in which different loss functions can make more sense for the applied forecaster but also for the user of a forecast such as a politician or the CEO of a company. For example, the recent recession demonstrated that a good forecast of a rather extreme event might be of special interest beyond that of minimizing an average squared error: banks could have taken earlier measures to shelter against the turmoil, governments could have started stimulus packages in time, and firms might have circumvented their strong increase in inventories.

As argued by van Dijk and Franses (2003), a weighted squared forecast error can be used to place more weight on unusual events when evaluating forecast models. Specifically, they propose to use the loss function

$$\mathcal{L}_{i,t}^w = w_t e_{i,t}^2, \tag{3}$$

where the weight w_t is specified as

1. $w_{\text{left},t} = 1 - \widehat{F}(y_t)$, where $F(\cdot)$ is the cumulative distribution function of y_t, to overweight the left tail of the distribution. This gives rise to a "recession" loss function.

2. $w_{\text{right},t} = \widehat{F}(y_t)$, to overweight the right tail of the distribution. This gives rise to a "boom" loss function.

Obviously, the weighted loss function (3) collapses to the standard loss function (1) when equal weights $w_t = 1$ are imposed. This gives rise to the conventional "uniform" loss function.

Using a weighted loss function complicates things only slightly. To evaluate a forecast model i over a forecast period $T+1$ to $T+P$ simply requires to calculate the weighted mean squared forecast error

$$\text{MSFE}_i = \frac{1}{P}\sum_{t=T+1}^{T+P} w_t e_{i,t}^2. \tag{4}$$

In order to compare, say, model i to a benchmark model 0, one calculates the weighted loss difference

$$d_{i,t} = \mathcal{L}_{0,t}^w - \mathcal{L}_{i,t}^w = w_t e_{0,t}^2 - w_t e_{i,t}^2 \tag{5}$$

and averages over the the forecast period

$$\overline{d}_i = \frac{1}{P}\sum_{t=T+1}^{T+P} d_{i,t} = \frac{1}{P}\sum_{t=T+1}^{T+P} w_t e_{0,t}^2 - \frac{1}{P}\sum_{t=T+1}^{T+P} w_t e_{i,t}^2 \tag{6}$$

In the remainder of this paper, we will use this weighted loss and analyze the forecast accuracy of different models (which in turn are based on different indicators) with respect to the different weighting schemes introduced above.

Figure 1 depicts the empirical cumulative density function of the target variable in our application, namely the growth rate of euro area industrial production. It demonstrates that observations smaller than -0.04 and larger than 0.04 receive a particularly high weight in the analysis of recessions and booms, respectively. The evolution of euro area industrial production and of the weight series is displayed in Figure 2. In the upper panel, the extreme fall in euro area industrial production during the winter of 2008/2009 catches the eye. Hence, this event also dominates the recession weights (lower panel). However, the recession in 2001/2002 receives almost the same weights. Therefore, our results are not solely driven by a single event. On the flip side, the boom weights are particularly high during the rapid expansion in 2000 and in the period of 2006 to 2008 (middle panel).

4 Forecast Accuracy Tests

To analyze whether empirical loss differences between two or more competing models are statistically significant, there is a large number of tests proposed in the literature, among which the pairwise test introduced by Diebold and Mariano (1995) seems to be the most influential and most widely used. Therefore, we also apply it to our setting. We augment our analysis with three further tests which are designed to account for additional important features of the forecast evaluation problem and which have not been used very often in applied work. First, the test proposed by Clark and West (2007) takes into account that our benchmark model—a simple AR(1) model—is nested in all the competing models to which early indicators are added. Second, the test suggested by Hansen (2005) circumvents the problem of data snooping that arises when a number of pairwise tests are conducted. Finally, the fluctuation test by Giacomini and Rossi (2008) is useful to examine whether the relative forecast performance of one model has changed over time relative to the benchmark. In the following, we briefly introduce these test.

4.1 Modified Diebold-Mariano Test

The standard way to discriminate between the forecasting performances of two competing models is to apply the forecast accuracy test proposed by Diebold and Mariano (1995). In this paper, we apply the modified Diebold-Mariano (MDM) test proposed by Harvey et al. (1997), which corrects for a small sample bias. It evaluates whether the average loss differences between the two models is significantly different from zero. Hence, it

is a pairwise test that is designed to compare two models at a time, say, model i with benchmark model 0. Specifically, the null hypothesis of the MDM test is that of equal forecast performance,

$$E\left[d_{i,t}\right] = E\left[\mathcal{L}_{0,t}^{w} - \mathcal{L}_{i,t}^{w}\right] = 0. \tag{7}$$

Following Harvey et al. (1997), we use the MDM test statistic

$$\text{MDM} = \left(\frac{P+1-2h+P^{-1}h(h-1)}{P}\right)^{1/2} \widehat{V}(\overline{d}_i)^{-1/2}\overline{d}_i, \tag{8}$$

where h is the forecast horizon and $\widehat{V}(\overline{d}_i)$ the estimated variance of series $d_{i,t}$. The MDM test statistic is compared with a critical value from the t-distribution with $P-1$ degrees of freedom.

4.2 Forecast accuracy test for nested models

In our setting presented in more detail below, then benchmark is an AR(1) model against which competing models augmented with more lags and additional indicators are tested. Hence, the benchmark model is nested in the competing models. When testing the null hypothesis of equal forecast accuracy for two nested models, a complication arises as argued by, inter alia, Clark and McCracken (2001) and Clark and West (2007). Consider the typical case in the applied forecast evaluation literature that a simple benchmark model is compared with a rival model which is augmented by additional explanatory variables such as further lags or indicators. Under the null, the additional variables are useless and their coefficients are zero. Estimating these coefficients introduces noise in the derived forecasts of the rival model. Hence, under the null, the forecast accuracy of the parsimonious benchmark model is higher than (and not equal to) that of the larger rival model. Neglecting this fact leads to undersized tests with poor power, see Clark and McCracken (2001) and Clark and West (2005). In this sense, conventional tests favor the parsimonious model too often. Therefore, Clark and West (2007) propose an adjusted test that takes the nested model structure into account.

Specifically, for a test in the spirit of Diebold and Mariano (1995), Clark and West (2007) define the adjustment term

$$\bar{a}_i = \frac{1}{P}\sum_{t=1}^{P} w_t \left(y_{0,t}^{f} - y_{i,t}^{f}\right)^2, \tag{9}$$

where $y_{0,t}^{f}$ is the forecast of the parsimonious benchmark model and $y_{i,t}^{f}$ is the forecast of the augmented rival model. As they consider an unweighted loss functions, they set $w_t = 1$.

The test statistic is defined as

$$\text{CW} = \widehat{V}(\overline{d}_i - \bar{a}_i)^{-1/2} \left(\overline{d}_i - \bar{a}_i\right), \tag{10}$$

where $\widehat{V}(\overline{d}_i - \bar{a}_i)$ is the estimated variance of the adjusted loss difference $\overline{d}_i - \bar{a}_i$. Note that it is essential that the forecasts be computed from a rolling regression. As demonstrated in a simulation study by Clark and West (2007), using forecasts computed from a rolling regression scheme and applying the normal distribution leads to a fairly good but somewhat undersized test. For example a test with 10 percent nominal size will typically have a true size between 5 and 10 percent. For our purpose, this should be a good approximation.

4.3 Superior Predictive Ability Test

Conventional econometric techniques for forecast evaluation focus on the comparison of two models at a time. Applying such pairwise tests sequentially to a number of models gives rise to the problems related to multiple testing procedures, particularly invalidating standard critical values. Effectively, comparing several different models to a benchmark model may result in spuriously identifying a superior model just by chance. To account for this data snooping problem we apply the test for superior predictive ability (SPA) proposed by Hansen (2005) which is based on the seminal paper by White (2000). The idea of this test is basically to compare a benchmark forecast model simultaneously to the whole set of m rival forecast models with the null hypothesis being that the benchmark is not inferior to any of the rivals. The null is formulated as the multiple hypothesis

$$H_0 : E(d_{i,t}) \leq 0 \quad \forall i = 1, \ldots, m. \tag{11}$$

and is rejected when at least one of the rival models yields significantly more accurate forecasts—and thus a smaller expected loss—than the benchmark model.

Of course, the expectation of $d_{i,t}$ is unknown, but it can be consistently estimated with the sample mean $\overline{d}_i$, $i = 1, ..., m$. White (2000) proposes the reality check test statistic

$$RC = \max_k P^{1/2} \overline{d}_i \,. \tag{12}$$

Note that the limiting distribution of RC is not unique under the null hypothesis. Therefore, the stationary bootstrap method of Politis and Romano (1994) is utilized.

As a major drawback, the RC test depends heavily on the set of competing models. If this set contains poor or irrelevant models delivering bad forecasts then the test is conservative in the sense that the critical value, which the RC statistic has to exceed in order to reject the null, increases with the number of included alternatives. Hence, adding enough

irrelevant models could, in principle, lead to accepting the null hypothesis no matter how good a single competing model might be. As a solution to this problem, Hansen (2005) proposes the studentized test statistic

$$SPA = \max\left[\max_k \widehat{V}(\overline{d}_i)^{-1/2}\overline{d}_i, 0\right], \tag{13}$$

where $\widehat{V}(\overline{d}_i)$ denotes the consistently estimated variance of $\overline{d}_i$. Assuming that irrelevant models deliver high forecast errors, the studentization downweights such models. Thereby, the size of the *SPA* test should be stable even if irrelevant models are added. Since the limiting distribution of the test statistic is not unique under the null hypothesis, a stationary bootstrap is used. Moreover, the distribution theory requires the use of a rolling estimation window in contrast to the recursive scheme used for the other tests.

4.4 Fluctuation Test

To analyze the stability of the forecasting performance over time, we implement the fluctuation test proposed by Giacomini and Rossi (2008). The test is based on the idea that due to potential structural instabilities—in our context possibly as the consequence of booms or recessions—the relative forecast performance of two competing models may change. Therefore, the authors propose to assess the development of a local loss difference over time in contrast to concentrating on the average (global) loss difference as in conventional tests. This may supply important information for a forecaster. In particular, indicator models that deliver accurate forecast only in specific situations or only at the beginning of the historical out-of-sample experiment might be downweighted.

To implement the fluctuation test, Giacomini and Rossi (2008) calculate the centered local loss differences of the Diebold-Mariano type,

$$\overline{d}_{i,t}^{\,\text{local}} = \frac{1}{Q}\sum_{\tau=t-Q/2}^{t+Q/2-1} \widehat{V}(\overline{d}_i)^{-1/2} d_{i,\tau}, \quad t = T+Q/2+1,\ldots,T+P-Q/2+1, \tag{14}$$

and check whether this sequence crosses the appropriate critical values which can be derived from a non-standard limiting distribution and are provided by the authors. If it does, then an instability is detected. Note that in our application below we calculate the forecasts from a rolling regression scheme.

When interpreting the results of the fluctuation test in comparison to a conventional Diebold-Mariano test, one should keep in mind that the null hypothesis of equal forecast accuracy is tested against slightly different alternatives. In the conventional approach, the alternative hypothesis is that one of the two models delivers a smaller expected loss than

the other *on average* over a fixed evaluation period. Hence, the approach presupposes structural invariance. In contrast, the fluctuation test uses the alternative hypothesis that one of the two models delivers a smaller expected loss at *some point* in the evaluation period. As this point is unknown, to prevent the test from spuriously detect instability, the absolute critical values are larger than in the conventional approach. This result is well known from standard structural break tests, such as the "sup" tests discussed by Andrews (1993). Therefore, in finite samples it might well be the case that the null hypothesis of equal forecast accuracy is rejected on average (assuming structural stability) but not locally (dropping the assumption of structural stability).

5 Empirical Setup

5.1 Database

We consider seven different business cycle indicators that are often used for the prediction of economic growth in the euro area. These indicators are constructed and published by different institutions such as the European Commission, the OECD, the ZEW, the DZ-Bank, and the CEPR. Table 1 contains a list of the indicators and their components. Our target series is the the year-over-year (yoy) growth rate of the industrial production index for the euro area as published by Eurostat. Although industrial production accounts only for one third of the total GDP, it is regarded as a well-suited and quickly available business cycle indicator as argued, inter alia, by Breitung and Jagodzinski (2001)). Our sample spans from 1992M01 to 2009M6.

5.2 Forecast model

In our forecast exercise we consider the standard autoregressive distributed lag (ADL) model for generating forecasts. The h-step-ahead model is given by

$$y_{t+h} = \alpha + \sum_{i=1}^{p} \phi_i y_{t+1-i} + \sum_{j=1}^{r} \theta_j x_{t+1-j} + \varepsilon_t \tag{15}$$

where y_t is the year-on-year growth rate of euro area industrial production and x_t denotes one of the aforementioned leading indicators which are taken as exogenous. Hence, we refrain from modeling feedback effects. We allow for a maximum of 12 lags both for the endogenous and the exogenous variable. The lag length is chosen via the AIC criterion. We employ a rolling forecasting scheme as required for the Hansen test. The initial estimation period ranges from 1992:01 to 1999:12 ($T = 96$) which is moved forward through up to

2009:05. At each point at time equation (15) is re-specified and before the forecasts are calculated. The initial forecast date is 2000:01 and the final forecast date is 2009:06 minus the forecast horizon. We generate short-term ($h = 1$), medium-term ($h = 6$) and long-term forecasts ($h = 12$). The number of calculated forecasts ranges from $P = 114$ for $h = 1$ to $P = 102$ for $h = 12$. We employ two benchmark models, an AR(1) model which is always nested in (15) and an AR(p) model.

6 Results

In a first step, we report the uniform, boom and recession weighted MSFE for all indicator models and the autoregressive benchmark models (Table 2). As a general result, the average forecast errors based on the uniform weighting scheme are strongly driven by the forecast errors made during recessions which are substantially higher than during booms. This holds for all models and forecast horizons. It implies that improvements in terms of indicator construction and model building should aim at better predictions of recession periods.

Comparing the indicators, we find that their ranking in some—but by far not in all—cases differs considerably between boom and recession periods. For the short-term forecasts ($h = 1$), we observe that the EJ indicator ranks as number 1 or 2 in all weighting schemes. Also, the EC indicator always ranks as number 3 or 4. Hence the relative performance of these indicators is unaffected by the specific economic situation. On the other hand, the relative performance of the ESI and OECD indicators depend on whether a boom or a recession has to be predicted. While the ESI is particularly useful in recessions, the OECD indicator has its strengths in booms. Overall, it is reassuring that all indicator models outperform the autoregressive benchmark models. The differences are particularly pronounced for recession forecasts.

Forecasting six months ahead leads to a somewhat different picture. Now the OECD indicator uniformly outperforms its competitors by a noticeable amount. The FAZ indicator follows closely behind for boom forecasts but is much less suited for recession forecasts. In contrast, the ZEW indicator works well for recession forecasts but ranks only as number 7 for boom forecasts. The EJ indicator which performed well for the short horizon cannot be recommended for the 6-month horizon.

Looking at the 12-month forecasts, the AR(1) model becomes number 2 on average and number 1 for boom forecasts. It is only outperformed by the FAZ indicator which works particularly well for recession forecasts. All the other indicators do not seem to add useful information to the simple autoregressive model.

In practice, the choice of an appropriate indicator should depend on both the forecast horizon and on the specific loss function. Forecasters who particularly dislike forecast errors during recessions should use a slightly different set of indicators than forecasters who are more interested in correct boom prediction. For example, at the 1-month horizon the top three models for recessions are based on the OECD, EJ, and FAZ indicators while the top three models for booms are based on the EJ, ESI, and EC indicators.

In a second step, the modified Diebold-Mariano test is used to check the significance of the above finding, see Tables 4 to 6. At the horizon of one month all indicator forecasts yield significantly smaller losses than the benchmark AR(1) model and the EJ indicator outperforms most of its competitors no matter which weighting scheme is used while some indicators are significantly better than other only in specific situations. For example, the FAZ indicator is significantly dominates 5 of its competitors during booms but not during recessions. At the horizon of six months the advantage of the OECD and FAZ indicators is corroborated by the Diebold-Mariano test, particularly for boom forecasts. For recession forecasts at the 6-month horizon, the EC, ZEW, FAZ, and OECD indicators are indistinguishable by the Diebold-Mariano test, even though the differences in MSFE between, e.g., the OECD and the FAZ indicator are considerable. At the horizon of 12 months no indicator is able to dominate the benchmark models, not even the FAZ indicator that has smaller MSFE albeit not significantly so. For all horizons and weighting schemes, the CFI exhibits a poor performance. A similar result holds for the ZEW indicator that is only useful for medium-term recession forecasts. However, we are careful with these test results because, as argued before, there a several caveats to take into account. Therefore, we supplement the Diebold-Mariano test and possibly qualify its results in the following.

The Clark-West test is computed to reassess the performance of the indicator models in comparison to the AR(1) benchmark. Since the modified Diebold-Mariano test is biased in favor of the nested AR(1) model, the results should be more in favor of the indicator models. Thus, for $h = 1$ the result is replicated that all indicator models outperform the benchmark. For $h = 6$, again all indicator models dominate the benchmark. This is different to the results of the Diebold-Mariano test that characterizes only few indicator models as significantly more accurate. At the 12-months horizon the Diebold-Mariano test does not find a single indicator model that outperforms the benchmark, while the Clark-West test identifies the FAZ indicator as being significantly better for the uniform and boom weighting schemes and almost significant (p-value of 0.12) for the recession weighting scheme. Note that during booms, also the EC, ZEW, and OECD indicators beat the AR(1) model. Overall, it pays off to use the Clark-West test.

The SPA test of Hansen is used to take into account that we are ultimately interested in

comparing each of the models simultaneously to all its competitors. Pairwise significance as attested by the Diebold-Mariano test might be spurious in some cases. In Table 7, we test for each model whether it is significantly outperformed by at least one of its competitors. For the AR(1) and AR(p) models, this is in fact the case at forecast horizons of $h = 1$ and $h = 6$, but not at $h = 12$. Also the CFI and ESI indicators are almost always dominated by at leat one competitor and may therefore be safely disregarded in forecasting exercises. The EJ indicator that was shown to perform excellent at the 1-month horizon is not of much use for medium and long-term forecasts as the null of equal predictive accuracy is rejected with p-values between 0.02 and 0.13. The ZEW indicator is a borderline case with p-values around 0.12 for $h = 1$ and 0.07 for $h = 6$. Its main strength seems to be the long-term forecast for which the null of equal predictive accuracy cannot be rejected at a safe margin. The remaining three indicators (EC, FAZ, OECD) are not significantly dominated by any competitor, irrespective of the forecast horizon or the weighting scheme.

Finally, we use the fluctuation tests to check the structural stability of the modified Diebold-Mariano (MDM) test results. In Figure 3, the MDM based fluctuation statistics for the horizon of $h = 1$ are displayed over the period from the beginning of 2005 to the middle of 2009 (note that the statistics are centered so that the last months of the sample cannot be considered). Each statistic refers to a pairwise test of the respective indicator model against the AR(1) benchmark. A value above the upper critical value indicates that the indicator is significantly more accurate than the benchmark while a value below the lower critical value indicates the opposite case. The variance of the statistics suggest that the superiority of the indicator models over the AR(1) benchmark ist not uniform over the whole forecast sample. In particular, average loss (using the uniform weights) of most the indicator models is statistically indistinguishable from that of the benchmark model during 2007 and most of 2008. However, during 2005/06 and since the end of 2008, the relative performance of some indicators is significantly better. These are periods of considerable changes in industrial production. This indicates that the simple benchmark might be better suited for rather tranquil times while the strength of the indicators is to contain early information on changes in business cycle. This impression is, however, only weakly supported by the medium-term and long-term fluctuation tests, where the fluctuation tests do not detect much instability, see Figures 4 and 5. This is mainly due to the fact that to signal local significance of the MDM test, a higher critical value has to be crossed than to signal average significance as reported by the standard MDM test discussed above. Nevertheless, within the interval of insignificance, we still observe that the test are more in favor of the simple benchmark during the period of 2007 and the first half of 2008.

7 Summary

In this paper we assessed the predictive abilities of seven widely recognized leading indicators for euro area industrial production. We went beyond standard forecast evaluation approaches in several respects, taking up recent methodological developments. We allowed for departures from the uniform symmetric quadratic loss function typically used in forecast evaluation exercises. Specifically, we overweighed forecast errors during periods of high or low growth rates to check whether how the indicators perform during boos and recessions, i.e., in times of particularly high demand for good forecasts. It turned out that some indicators are well-suited for booms or recessions only while others are largely unaffected by the business cycle situation.

We also took the issue of nested models into account when comparing indicator models with a simple autoregressive benchmark. Unlike the standard Diebold-Mariano test, the test proposed by Clark and West (2007) identified all indicators as significantly outperforming the benchmark at short to medium-term forecast horizons. This result confirms the usefulness of the seven early indicators for euro area industrial production.

In order to prevent the problem of data snooping when searching for the best of the seven indicators by performing multiple pairwise tests, we implemented the test for superior predictive ability proposed by Hansen (2005). The results pointed to the existence of a group of three top indicators (EC, FAZ, OECD) that are generally not dominated by others. However, it is not possible to significantly discriminate between these three. For short-term forecasts, also the Business Climate Indicator (EJ) published by the European Commission performed excellent.

Finally, we implemented the fluctuation test introduced by Giacomini and Rossi (2008) to assess the forecasting stability of each model both on average and during booms and recessions. It indicated that the simple autoregressive benchmark model might be difficult to beat in rather tranquil times while the strength of the indicators is to contain early information on booms and recessions.

References

Andrews, Donald W. (1993), "Tests for parameter instability and structural change with unknown change point," *Econometrica*, 61, 821–856.

Bodo, G., R. Golinelli, and G. Parigi (2000), "Forecasting industrial production in the euro area," *Empirical economics*, 25(4), 541–561.

Breitung, Jörg, and Doris Jagodzinski (2001), "Prognoseeigenschaften alternativer Indikatoren für die Konjunkturentwicklung in Deutschland," *Konjunkturpolitik*, 47, 292–314.

Caggiano, G., G. Kapetanios, and V. Labhard (2009), "Are more data always better for factor analysis? Results for the euro area, the six largest euro area countries and the UK," Working paper series, European Central Bank.

Clark, T.E., and M.W. McCracken (2009), "Averaging forecasts from VARs with uncertain instabilities," *Journal of Applied Econometrics*, 25(1), 5–21.

Clark, Todd E., and Michael W. McCracken (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85–110.

Clark, Todd E., and Kenneth D. West (2005), "Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis," *Journal of Econometrics*, 135, 155–186.

Clark, Todd E., and Kenneth D. West (2007), "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Econometrics*, 138, 291–311.

Diebold, Francis X., and Roberto S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13(3), 253–263.

Fichtner, F., R. Ruffer, and B. Schnatz (2009), "Leading indicators in a globalised world," Working paper series, European Central Bank.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2003), "Do financial variables help forecasting inflation and real activity in the euro area?" *Journal of Monetary Economics*, 50(6), 1243–1255.

Giacomini, Raffaella, and Barbara Rossi (2008), "Forecasting Comparisons in Unstable Environments," Working Paper 08–04, Duke University, Department of Economics.

Hansen, Peter Reinhard (2005), "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics*, 23(4), 365–380.

Harvey, David I., Stephen J. Leybourne, and Paul Newbold (1997), "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 13, 281–291.

Marcellino, Massimiliano (2008), "A linear benchmark for forecasting GDP growth and inflation?" *Journal of Forecasting*, 27(4), 305–340.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson (2003), "Macroeconomic forecasting in the Euro area: Country specific versus area-wide information," *European Economic Review*, 47(1), 1–18.

Milas, Costas, and Philip Rothman (2008), "Out-of-sample forecasting of unemployment rates with pooled STVECM forecasts," *International Journal of Forecasting*, 24(1), 101–121.

Politis, Dimitris N., and Joseph P. Romano (1994), "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89(428), 1303–1313.

Rossi, B., and T. Sekhposyan (2010), "Have economic models' forecasting performance for US output growth and inflation changed over time, and when?" *International Journal of Forecasting*, In press.

van Dijk, Dick, and Philip Hans Franses (2003), "Selecting a Nonlinear Time Series Model using Weighted Tests of Equal Forecast Accuracy," *Oxford Bulletin of Economics and Statistics*, 65, 727–744.

White, Halbert (2000), "A reality check for data snooping," *Econometrica*, 68(5), 1097–1126.

Table 1: Overview over the euro area indicators

Indicator	**Components**	**Source**
European Sentiment Indicator (ESI)	Industry Confidence Indicator, Services Confindence Indicator Consumer Confidence Indicator (CFI) Construction Confidence Indicator Retail Trade Confidence Indicator	European Commission
Consumer Confidence Indicator (CFI)	Consumer surveys	European Commission
Business Climate Indicator (EJ)	Industry survey about: production trends in recent months, order books export order books, stocks and production expectations	European Commission
FAZ-Euro-Indicator (FAZ)	New job vacancies, order entries, Reuter purchasing manager´s index (PMI), building and planning permissions, production, interest rate spread, cosumer confidence, Morgan-Stanley- Capital-International Index, real money (M3)	DZ-Bank
OECD Composite Indicator (OECD)	Composite by individual OECD indicators for EU-12: variables for surveys by national institutes, new job vacancies, orders inflow/demand, spread of interest rates, production, finished goods stocks, passenger car registration, other national indicators	Organisation for Economic Co-operation and Development (OECD)
ZEW Indicator of Economic Sentiment (ZEW)	Medium-term expectations for development of the macroeconomic trend, inflation rate, short-term and long-term interest rates, stockmarket, exchange rates, profit situation of different German industries (only financial experts)	Centre for European Economic Research (ZEW)
EuroCoin (EC)	Data from 11 categories: industrial production, producer prices, monetary aggregates, interest rates, financial variables, exchange rates, surveys by the European Commission, surveys by national institutes, external trade, labour market	Centre for Economic Policy Research (CEPR)

Table 2: Root Mean Squared Forecast Errors

	Uniform		Boom		Recession	
	MSE	Rank	MSE	Rank	MSE	Rank
			$h = 1$			
AR(1)	0.022	9	0.014	9	0.027	9
AR	0.020	8	0.013	8	0.025	8
ESI	0.015	4	0.012	5	0.017	2
EJ	0.013	1	0.011	2	0.015	1
CFI	0.015	5	0.012	6	0.017	5
EC	0.014	3	0.012	4	0.017	3
ZEW	0.016	7	0.013	7	0.018	6
FAZ	0.016	6	0.011	3	0.019	7
OECD	0.014	2	0.011	1	0.017	4
			$h = 6$			
AR(1)	0.053	8	0.023	5	0.072	9
AR	0.050	7	0.023	6	0.066	7
ESI	0.049	6	0.025	8	0.065	6
EJ	0.046	5	0.022	4	0.062	5
CFI	0.054	9	0.028	9	0.070	8
EC	0.041	3	0.020	3	0.054	3
ZEW	0.042	4	0.024	7	0.054	2
FAZ	0.041	2	0.018	2	0.055	4
OECD	0.034	1	0.016	1	0.045	1
			$h = 12$			
AR(1)	0.063	2	0.027	1	0.084	2
AR	0.065	5	0.032	5	0.087	5
ESI	0.071	8	0.037	7	0.093	9
EJ	0.067	7	0.038	8	0.087	6
CFI	0.072	9	0.044	9	0.092	8
EC	0.064	3	0.030	3	0.085	3
ZEW	0.065	4	0.034	6	0.086	4
FAZ	0.058	1	0.030	2	0.076	1
OECD	0.065	6	0.030	4	0.087	7

Notes: This Table reports the root MSFEs and the corresponding ranking for each forecasting horizon and weighting scheme.

Table 3: Results of the Clark-West Test

	Total			Boom			Recession		
	1	6	12	1	6	12	1	6	12
AR	0.000	0.041	0.170	0.000	0.002	0.157	0.015	0.083	0.459
ESI	0.000	0.015	0.496	0.000	0.000	0.297	0.001	0.039	0.260
EJ	0.000	0.025	0.201	0.000	0.000	0.378	0.001	0.042	0.233
CFI	0.001	0.007	0.458	0.000	0.003	0.294	0.002	0.038	0.462
EC	0.000	0.022	0.130	0.000	0.000	0.016	0.000	0.034	0.316
ZEW	0.001	0.039	0.104	0.000	0.004	0.077	0.005	0.047	0.294
FAZ	0.000	0.014	0.079	0.000	0.000	0.008	0.001	0.026	0.117
OECD	0.000	0.040	0.156	0.000	0.000	0.007	0.000	0.052	0.431

Notes: Table reports p-values for the two-sided modified Clark-West test. A p-value smaller than 0.05 indicates that the row indicator has a significantly smaller MSE than the nested AR(1) benchmark model.

Table 4: Modified Diebold-Mariano test for uniform weights ($w_t = 1$)

$h=1$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)		−1.66 (0.050)	−2.73 (0.004)	−2.97 (0.002)	−2.45 (0.008)	−3.16 (0.001)	−2.23 (0.014)	−2.72 (0.004)	−2.99 (0.002)	0	8
AR	1.66 (0.050)		−2.53 (0.006)	−2.85 (0.003)	−2.27 (0.013)	−3.03 (0.001)	−2.03 (0.023)	−2.66 (0.004)	−2.94 (0.002)	1	7
ESI	2.73 (0.004)	2.53 (0.006)		−1.83 (0.035)	0.74 (0.230)	−0.10 (0.460)	0.86 (0.196)	0.82 (0.206)	−0.26 (0.398)	2	1
EJ	2.97 (0.002)	2.85 (0.003)	1.83 (0.035)		2.39 (0.009)	1.11 (0.136)	1.85 (0.034)	1.66 (0.050)	0.91 (0.183)	6	0
CFI	2.45 (0.008)	2.27 (0.013)	−0.74 (0.230)	−2.39 (0.009)		−0.63 (0.265)	0.53 (0.298)	0.41 (0.343)	−0.77 (0.222)	2	1
EC	3.16 (0.001)	3.03 (0.001)	0.10 (0.460)	−1.11 (0.136)	0.63 (0.265)		0.93 (0.177)	1.25 (0.107)	−0.24 (0.407)	2	0
ZEW	2.22 (0.014)	2.03 (0.023)	−0.86 (0.196)	−1.85 (0.034)	−0.53 (0.298)	−0.93 (0.177)		−0.14 (0.445)	−0.99 (0.161)	2	1
FAZ	2.72 (0.004)	2.66 (0.004)	−0.82 (0.206)	−1.66 (0.050)	−0.41 (0.343)	−1.25 (0.107)	0.14 (0.445)		−1.12 (0.132)	2	1
OECD	2.99 (0.002)	2.94 (0.002)	0.26 (0.398)	−0.91 (0.183)	0.77 (0.222)	0.24 (0.407)	0.99 (0.161)	1.12 (0.132)		2	0

$h=6$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)		−0.90 (0.186)	−0.81 (0.209)	−1.24 (0.108)	0.08 (0.467)	−1.48 (0.071)	−1.22 (0.112)	−1.57 (0.059)	−1.45 (0.075)	0	3
AR	0.90 (0.186)		−0.12 (0.454)	−1.50 (0.068)	1.61 (0.055)	−1.68 (0.048)	−1.32 (0.094)	−1.84 (0.034)	−1.53 (0.064)	1	5
ESI	0.81 (0.209)	0.12 (0.454)		−1.42 (0.079)	1.74 (0.042)	−1.88 (0.031)	−1.45 (0.075)	−2.16 (0.017)	−1.65 (0.051)	1	5
EJ	1.24 (0.108)	1.50 (0.068)	1.42 (0.079)		1.76 (0.040)	−1.63 (0.053)	−1.09 (0.139)	−1.78 (0.039)	−1.45 (0.075)	3	3
CFI	−0.08 (0.467)	−1.61 (0.055)	−1.74 (0.042)	−1.76 (0.040)		−1.93 (0.028)	−1.59 (0.057)	−2.11 (0.019)	−1.75 (0.042)	0	7
EC	1.48 (0.071)	1.68 (0.048)	1.88 (0.031)	1.63 (0.053)	1.93 (0.028)		0.57 (0.285)	−0.04 (0.486)	−1.28 (0.101)	5	0
ZEW	1.22 (0.112)	1.32 (0.094)	1.45 (0.075)	1.09 (0.139)	1.59 (0.057)	−0.57 (0.285)		−0.46 (0.322)	−1.58 (0.058)	3	1
FAZ	1.57 (0.059)	1.84 (0.034)	2.16 (0.017)	1.78 (0.039)	2.11 (0.019)	0.04 (0.486)	0.46 (0.322)		−1.13 (0.131)	5	0
OECD	1.45 (0.075)	1.53 (0.064)	1.65 (0.051)	1.45 (0.075)	1.75 (0.042)	1.28 (0.101)	1.58 (0.058)	1.13 (0.131)		6	0

$h=12$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)		0.49 (0.312)	1.11 (0.136)	0.96 (0.169)	1.52 (0.066)	0.38 (0.352)	0.49 (0.312)	−0.57 (0.286)	0.56 (0.287)	1	0
AR	−0.49 (0.312)		1.45 (0.075)	0.61 (0.272)	1.72 (0.044)	−0.43 (0.335)	−0.06 (0.476)	−1.05 (0.148)	0.03 (0.488)	2	0
ESI	−1.11 (0.136)	−1.45 (0.075)		−0.85 (0.198)	0.50 (0.310)	−1.56 (0.061)	−1.52 (0.066)	−1.32 (0.095)	−1.13 (0.131)	0	4
EJ	−0.96 (0.169)	−0.61 (0.272)	0.85 (0.198)		1.11 (0.136)	−1.43 (0.079)	−0.86 (0.196)	−1.25 (0.106)	−0.49 (0.313)	0	1
CFI	−1.52 (0.066)	−1.72 (0.044)	−0.50 (0.310)	−1.11 (0.136)		−1.93 (0.028)	−1.72 (0.044)	−1.67 (0.049)	−1.33 (0.093)	0	6
EC	−0.38 (0.352)	0.43 (0.335)	1.56 (0.061)	1.43 (0.079)	1.93 (0.028)		0.49 (0.311)	−0.80 (0.213)	0.42 (0.337)	3	0
ZEW	−0.49 (0.312)	0.06 (0.476)	1.52 (0.066)	0.86 (0.196)	1.72 (0.044)	−0.49 (0.311)		−1.08 (0.142)	0.05 (0.479)	2	0
FAZ	0.57 (0.286)	1.05 (0.148)	1.32 (0.095)	1.25 (0.106)	1.67 (0.049)	0.80 (0.213)	1.08 (0.142)		0.82 (0.207)	2	0
OECD	−0.56 (0.287)	−0.03 (0.488)	1.13 (0.131)	0.49 (0.313)	1.33 (0.093)	−0.42 (0.337)	−0.05 (0.479)	−0.82 (0.207)		1	0

Notes: For each pair of models the modified DM test statistic is reported together with the two-sided p-value in brackets below. A negative sign indicates that the MSFE of row model is smaller than that of the column model and vice versa. The last two columns count the number of times the row model significantly outperforms its competitors (column "+") and are outperformed by its competitors (column "-").

Table 5: Modified Diebold-Mariano test for boom weights (w_{right})

$h=1$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)		−2.257 (0.013)	−2.311 (0.011)	−3.496 (0.000)	−2.143 (0.017)	−2.590 (0.005)	−1.735 (0.043)	−2.990 (0.002)	−3.901 (0.000)	0	8
AR	2.257 (0.013)		−1.021 (0.155)	−2.291 (0.012)	−0.690 (0.246)	−1.280 (0.102)	−0.402 (0.344)	−1.780 (0.039)	−2.633 (0.005)	1	3
ESI	2.311 (0.011)	1.021 (0.155)		−1.401 (0.082)	0.485 (0.314)	−0.575 (0.283)	0.640 (0.262)	−1.530 (0.064)	−1.511 (0.067)	1	3
EJ	3.496 (0.000)	2.291 (0.012)	1.401 (0.082)		1.821 (0.036)	0.811 (0.209)	1.567 (0.060)	0.319 (0.375)	−0.455 (0.325)	5	0
CFI	2.143 (0.017)	0.690 (0.246)	−0.485 (0.314)	−1.821 (0.036)		−0.846 (0.200)	0.209 (0.418)	−1.362 (0.088)	−1.968 (0.026)	1	3
EC	2.590 (0.005)	1.280 (0.102)	0.575 (0.283)	−0.811 (0.209)	0.846 (0.200)		0.964 (0.169)	−0.688 (0.246)	−1.050 (0.148)	1	0
ZEW	1.735 (0.043)	0.402 (0.344)	−0.640 (0.262)	−1.567 (0.060)	−0.209 (0.418)	−0.964 (0.169)		−1.415 (0.080)	−1.828 (0.035)	1	3
FAZ	2.990 (0.002)	1.780 (0.039)	1.530 (0.064)	−0.319 (0.375)	1.362 (0.088)	0.688 (0.246)	1.415 (0.080)		−0.586 (0.279)	5	0
OECD	3.901 (0.000)	2.633 (0.005)	1.511 (0.067)	0.455 (0.325)	1.968 (0.026)	1.050 (0.148)	1.828 (0.035)	0.586 (0.279)		5	0

$h=6$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)		0.080 (0.468)	0.774 (0.220)	−0.166 (0.434)	1.301 (0.098)	−1.095 (0.138)	0.528 (0.299)	−1.608 (0.055)	−2.442 (0.008)	1	2
AR	−0.080 (0.468)		1.724 (0.044)	−0.345 (0.365)	1.844 (0.034)	−1.077 (0.142)	0.565 (0.287)	−1.822 (0.036)	−2.300 (0.012)	2	2
ESI	−0.774 (0.220)	−1.724 (0.044)		−1.421 (0.079)	1.509 (0.067)	−1.834 (0.035)	−0.282 (0.389)	−2.803 (0.003)	−2.712 (0.004)	1	5
EJ	0.166 (0.434)	0.345 (0.365)	1.421 (0.079)		1.790 (0.038)	−1.308 (0.097)	0.879 (0.191)	−1.887 (0.031)	−2.767 (0.003)	2	3
CFI	−1.301 (0.098)	−1.844 (0.034)	−1.509 (0.067)	−1.790 (0.038)		−2.159 (0.017)	−0.936 (0.176)	−3.196 (0.001)	−2.749 (0.004)	0	7
EC	1.095 (0.138)	1.077 (0.142)	1.834 (0.035)	1.308 (0.097)	2.159 (0.017)		1.534 (0.064)	−1.186 (0.119)	−2.435 (0.008)	4	1
ZEW	−0.528 (0.299)	−0.565 (0.287)	0.282 (0.389)	−0.879 (0.191)	0.936 (0.176)	−1.534 (0.064)		−1.813 (0.036)	−2.592 (0.005)	0	3
FAZ	1.608 (0.055)	1.822 (0.036)	2.803 (0.003)	1.887 (0.031)	3.196 (0.001)	1.186 (0.119)	1.813 (0.036)		−1.100 (0.137)	6	0
OECD	2.442 (0.008)	2.300 (0.012)	2.712 (0.004)	2.767 (0.003)	2.749 (0.004)	2.435 (0.008)	2.592 (0.005)	1.100 (0.137)		7	0

$h=12$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)		0.798 (0.213)	1.400 (0.082)	1.656 (0.050)	1.692 (0.047)	0.701 (0.242)	1.048 (0.148)	0.527 (0.300)	0.511 (0.305)	3	0
AR	−0.798 (0.213)		2.286 (0.012)	1.116 (0.133)	1.301 (0.098)	−0.612 (0.271)	0.525 (0.300)	−0.550 (0.292)	−0.546 (0.293)	2	0
ESI	−1.400 (0.082)	−2.286 (0.012)		0.227 (0.411)	0.835 (0.203)	−1.743 (0.042)	−1.404 (0.082)	−1.750 (0.042)	−1.822 (0.036)	0	6
EJ	−1.656 (0.050)	−1.116 (0.133)	−0.227 (0.411)		0.533 (0.298)	−1.596 (0.057)	−1.048 (0.148)	−1.395 (0.083)	−1.495 (0.069)	0	4
CFI	−1.692 (0.047)	−1.301 (0.098)	−0.835 (0.203)	−0.533 (0.298)		−1.480 (0.071)	−1.107 (0.136)	−1.633 (0.053)	−1.419 (0.080)	0	5
EC	−0.701 (0.242)	0.612 (0.271)	1.743 (0.042)	1.596 (0.057)	1.480 (0.071)		1.103 (0.136)	−0.039 (0.484)	0.108 (0.457)	3	0
ZEW	−1.048 (0.148)	−0.525 (0.300)	1.404 (0.082)	1.048 (0.148)	1.107 (0.136)	−1.103 (0.136)		−1.066 (0.145)	−0.949 (0.172)	1	0
FAZ	−0.527 (0.300)	0.550 (0.292)	1.750 (0.042)	1.395 (0.083)	1.633 (0.053)	0.039 (0.484)	1.066 (0.145)		0.143 (0.443)	3	0
OECD	−0.511 (0.305)	0.546 (0.293)	1.822 (0.036)	1.495 (0.069)	1.419 (0.080)	−0.108 (0.457)	0.949 (0.172)	−0.143 (0.443)		3	0

Notes: See notes in Table 4.

Table 6: Modified Diebold-Mariano test for recession weights (w_{left})

$h=1$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)		−1.271 (0.103)	−2.563 (0.006)	−2.693 (0.004)	−2.323 (0.011)	−2.951 (0.002)	−2.098 (0.019)	−2.406 (0.009)	−2.646 (0.005)	0	7
AR	1.271 (0.103)		−2.557 (0.006)	−2.738 (0.004)	−2.330 (0.011)	−3.055 (0.001)	−2.106 (0.019)	−2.543 (0.006)	−2.749 (0.003)	0	7
ESI	2.563 (0.006)	2.557 (0.006)		−1.615 (0.055)	0.680 (0.249)	0.036 (0.486)	0.770 (0.222)	1.138 (0.129)	0.282 (0.389)	2	1
EJ	2.693 (0.004)	2.738 (0.004)	1.615 (0.055)		2.059 (0.021)	0.968 (0.168)	1.526 (0.065)	1.749 (0.042)	1.189 (0.118)	6	0
CFI	2.323 (0.011)	2.330 (0.011)	−0.680 (0.249)	−2.059 (0.021)		−0.428 (0.335)	0.529 (0.299)	0.837 (0.202)	−0.176 (0.430)	2	1
EC	2.951 (0.002)	3.055 (0.001)	−0.036 (0.486)	−0.968 (0.168)	0.428 (0.335)		0.757 (0.225)	1.648 (0.051)	0.422 (0.337)	3	0
ZEW	2.098 (0.019)	2.106 (0.019)	−0.770 (0.222)	−1.526 (0.065)	−0.529 (0.299)	−0.757 (0.225)		0.366 (0.358)	−0.570 (0.285)	2	1
FAZ	2.406 (0.009)	2.543 (0.006)	−1.138 (0.129)	−1.749 (0.042)	−0.837 (0.202)	−1.648 (0.051)	−0.366 (0.358)		−1.124 (0.132)	2	2
OECD	2.646 (0.005)	2.749 (0.003)	−0.282 (0.389)	−1.189 (0.118)	0.176 (0.430)	−0.422 (0.337)	0.570 (0.285)	1.124 (0.132)		2	0

$h=6$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)		−0.990 (0.162)	−1.054 (0.147)	−1.282 (0.101)	−0.401 (0.345)	−1.458 (0.074)	−1.329 (0.093)	−1.498 (0.069)	−1.369 (0.087)	0	4
AR	0.990 (0.162)		−0.664 (0.254)	−1.593 (0.057)	1.180 (0.120)	−1.644 (0.052)	−1.480 (0.071)	−1.702 (0.046)	−1.415 (0.080)	0	5
ESI	1.054 (0.147)	0.664 (0.254)		−1.230 (0.111)	1.508 (0.067)	−1.699 (0.046)	−1.486 (0.070)	−1.828 (0.035)	−1.440 (0.076)	1	4
EJ	1.282 (0.101)	1.593 (0.057)	1.230 (0.111)		1.492 (0.069)	−1.557 (0.061)	−1.326 (0.094)	−1.556 (0.061)	−1.308 (0.097)	2	4
CFI	0.401 (0.345)	−1.180 (0.120)	−1.508 (0.067)	−1.492 (0.069)		−1.688 (0.047)	−1.518 (0.066)	−1.763 (0.040)	−1.512 (0.067)	0	6
EC	1.458 (0.074)	1.644 (0.052)	1.699 (0.046)	1.557 (0.061)	1.688 (0.047)		−0.071 (0.472)	0.389 (0.349)	−1.121 (0.132)	5	0
ZEW	1.329 (0.093)	1.480 (0.071)	1.486 (0.070)	1.326 (0.094)	1.518 (0.066)	0.071 (0.472)		0.279 (0.390)	−1.182 (0.120)	5	0
FAZ	1.498 (0.069)	1.702 (0.046)	1.828 (0.035)	1.556 (0.061)	1.763 (0.040)	−0.389 (0.349)	−0.279 (0.390)		−1.071 (0.143)	5	0
OECD	1.369 (0.087)	1.415 (0.080)	1.440 (0.076)	1.308 (0.097)	1.512 (0.067)	1.121 (0.132)	1.182 (0.120)	1.071 (0.143)		5	0

$h=12$	AR(1)	AR	ESI	EJ	CFI	EC	ZEW	FAZ	OECD	+	-
AR(1)		0.370 (0.356)	0.965 (0.169)	0.473 (0.318)	1.126 (0.131)	0.249 (0.402)	0.264 (0.396)	−0.717 (0.238)	0.538 (0.296)	0	0
AR	−0.370 (0.356)		1.155 (0.125)	0.026 (0.490)	1.433 (0.077)	−0.338 (0.368)	−0.517 (0.303)	−1.043 (0.150)	0.183 (0.428)	1	0
ESI	−0.965 (0.169)	−1.155 (0.125)		−1.223 (0.112)	−0.178 (0.430)	−1.384 (0.085)	−1.362 (0.088)	−1.173 (0.122)	−0.860 (0.196)	0	2
EJ	−0.473 (0.318)	−0.026 (0.490)	1.223 (0.112)		1.509 (0.067)	−0.765 (0.223)	−0.401 (0.345)	−0.988 (0.163)	0.185 (0.427)	1	0
CFI	−1.126 (0.131)	−1.433 (0.077)	0.178 (0.430)	−1.509 (0.067)		−1.761 (0.041)	−1.736 (0.043)	−1.295 (0.099)	−0.889 (0.188)	0	5
EC	−0.249 (0.402)	0.338 (0.368)	1.384 (0.085)	0.765 (0.223)	1.761 (0.041)		0.181 (0.428)	−0.834 (0.203)	0.448 (0.327)	2	0
ZEW	−0.264 (0.396)	0.517 (0.303)	1.362 (0.088)	0.401 (0.345)	1.736 (0.043)	−0.181 (0.428)		−0.988 (0.163)	0.326 (0.372)	2	0
FAZ	0.717 (0.238)	1.043 (0.150)	1.173 (0.122)	0.988 (0.163)	1.295 (0.099)	0.834 (0.203)	0.988 (0.163)		0.847 (0.200)	1	0
OECD	−0.538 (0.296)	−0.183 (0.428)	0.860 (0.196)	−0.185 (0.427)	0.889 (0.188)	−0.448 (0.327)	−0.326 (0.372)	−0.847 (0.200)		0	0

Notes: See notes in Table 4.

Table 7: Results of the Hansen Test for Superior Predictive Ability (p-values)

	$h=1$			$h=6$			$h=12$		
	uniform	boom	recess.	uniform	boom	recess.	uniform	boom	recess.
AR(1)	0.03	0.03	0.03	0.03	0.03	0.03	0.41	0.56	0.36
AR	0.04	0.04	0.04	0.03	0.03	0.02	0.17	0.26	0.17
CFI	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.03
EC	0.34	0.61	0.31	0.10	0.13	0.10	0.25	0.44	0.23
EJ	0.70	0.99	0.70	0.03	0.03	0.02	0.11	0.13	0.10
ESI	0.11	0.14	0.09	0.01	0.01	0.01	0.08	0.08	0.08
FAZ	0.23	0.31	0.22	0.14	0.23	0.14	0.75	0.93	0.58
OECD	0.39	0.57	0.31	0.60	0.92	0.60	0.15	0.18	0.11
ZEW	0.12	0.13	0.12	0.07	0.09	0.07	0.17	0.27	0.17

Notes: Reported are p-values of SPA tests with the null hypothesis that the row model has equal predictive ability as all its competitor models against the alternative that at least one competitor yields more accurate predictions.

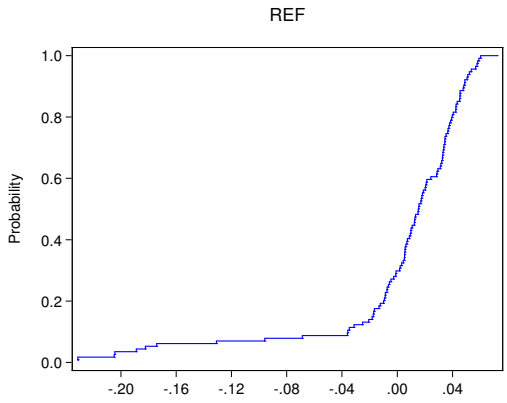

Figure 1: Empirical Cumulative Distribution Function $\widehat{F}(y_t)$

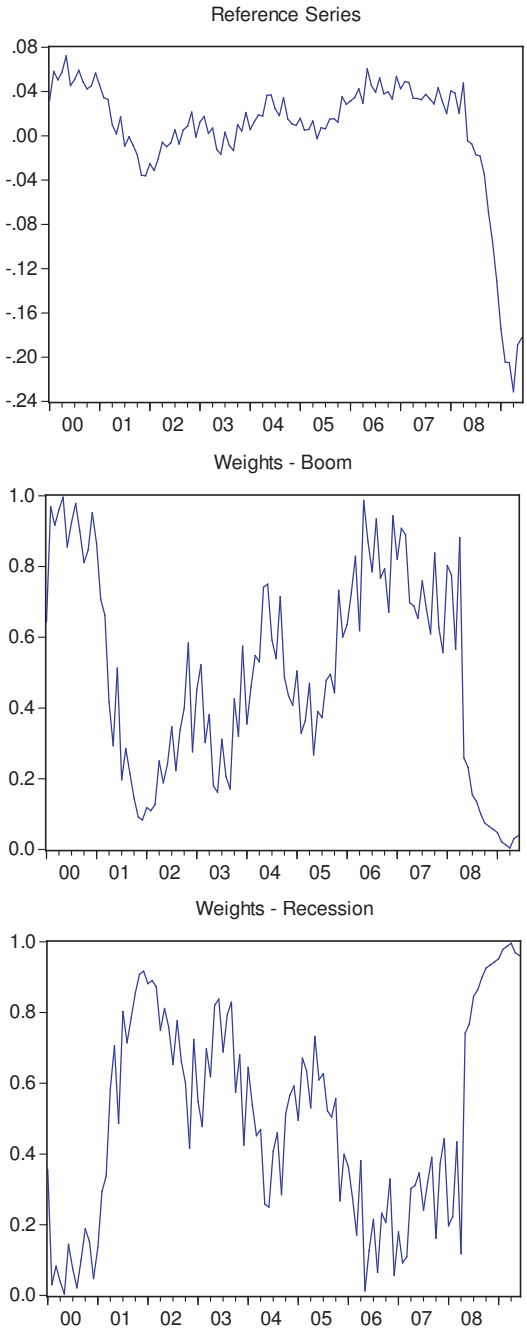

Figure 2: Reference Series and Weights

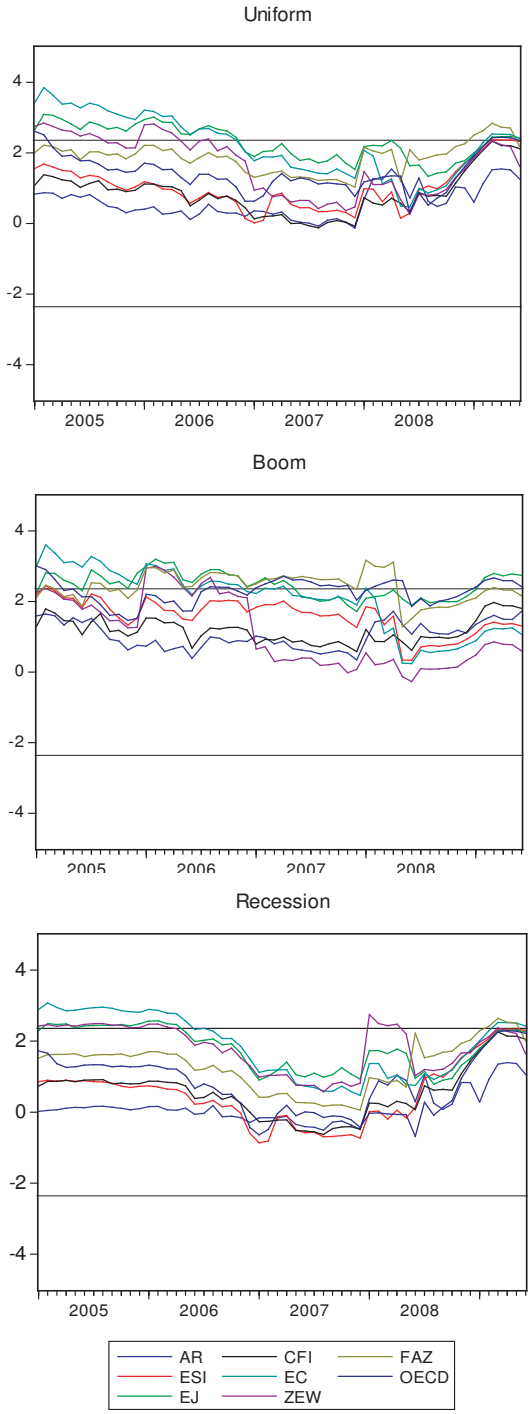

Figure 3: Fluctuation MDM test for $h = 1$ against the AR(1) benchmark

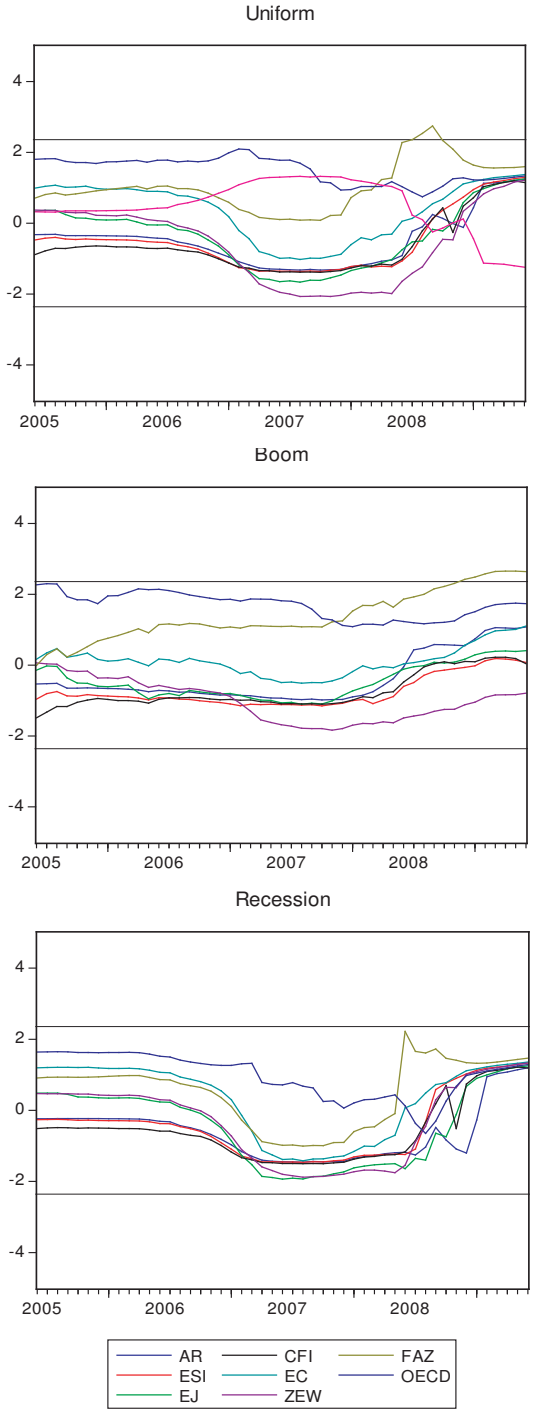

Figure 4: Fluctuation MDM test for $h = 6$ against the AR(1) benchmark

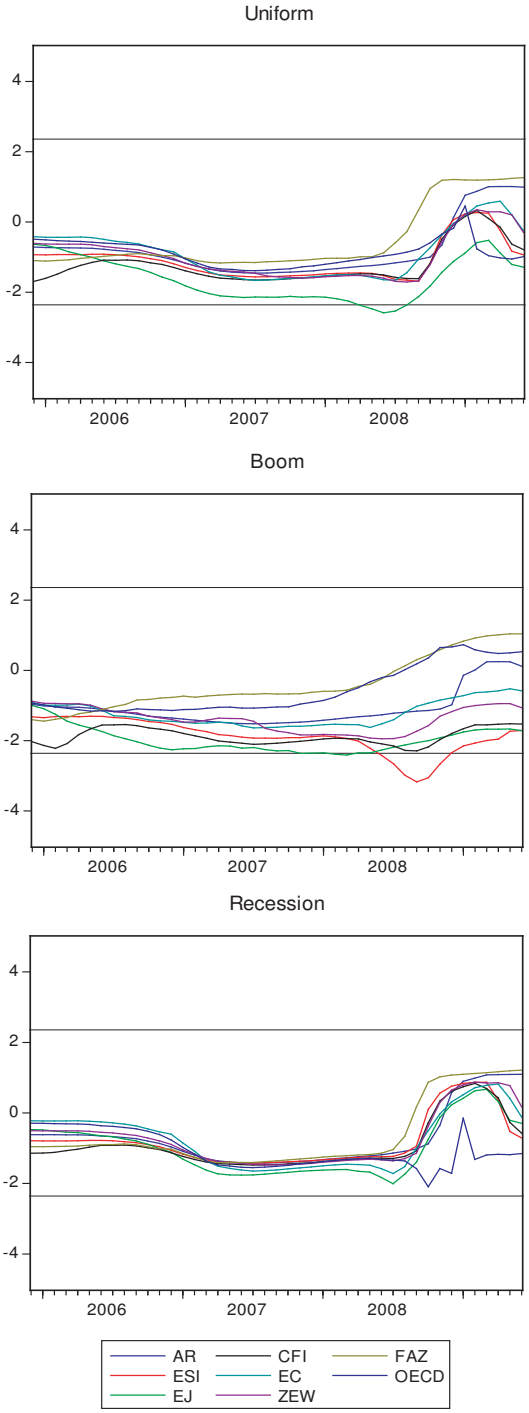

Figure 5: Fluctuation MDM test for $h = 12$ against the AR(1) benchmark